OPEN ACCESS

University of BRISTOL

Panphattarasap, P., & Calway, A. (2017). Visual place recognition using landmark distribution descriptors. In *Computer Vision - ACCV 2016: 13th Asian Conference on Computer Vision, ACCV 2016, Revised Selected Papers* (Vol. 10114 LNCS, pp. 487-502). (Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Vol. 10114 LNCS). Springer-Verlag Berlin. https://doi.org/10.1007/978-3-319-54190-7_30

Peer reviewed version

Link to published version (if available):
10.1007/978-3-319-54190-7_30

Link to publication record in Explore Bristol Research
PDF-document

This is the author accepted manuscript (AAM). The final published version (version of record) is available online via SpringerLink at https://link.springer.com/chapter/10.1007%2F978-3-319-54190-7_30 . Please refer to any applicable terms of use of the publisher.

University of Bristol - Explore Bristol Research
General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available:
http://www.bristol.ac.uk/red/research-policy/pure/user-guides/ebr-terms/

Visual place recognition using landmark distribution descriptors

Pilailuck Panphattarasap and Andrew Calway

Department of Computer Science
University of Bristol, UK

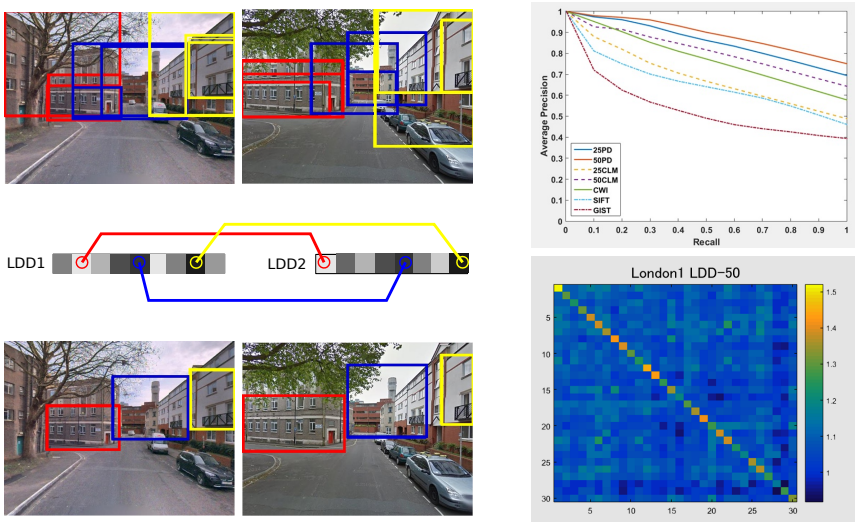
Abstract. Recent work by Sünderhauf et al. [1] demonstrated improved visual place recognition using proposal regions coupled with features from convolutional neural networks (CNN) to match landmarks between views. In this work we extend the approach by introducing descriptors built from landmark features which also encode the spatial distribution of the landmarks within a view. Matching descriptors then enforces consistency of the relative positions of landmarks between views. This has a significant impact on performance. For example, in experiments on 10 image-pair datasets, each consisting of 200 urban locations with significant differences in viewing positions and conditions, we recorded average precision of around 70% (at 100% recall), compared with 58% obtained using whole image CNN features and 50% for the method in [1].

1 Introduction

Visual place recognition is the task of matching a view of a place with a different view of the same place taken at a different time. If incorporated into a mapping framework, such as a topological representation of places, for example, then reliable and fast visual place recognition opens up the possibility of truly autonomous navigation, with applications in robotics and related areas. It would negate the need for a positioning infrastructure such as GPS and perhaps more interestingly, in respect of human-robot interaction, be more akin to the wayfinding techniques employed by humans.

Automated recognition of places based on visual information is however very challenging. It is highly dependent on the characteristics of places, the viewing positions and directions, and the environmental conditions in terms of light and visibility. Perspective effects, occlusions, changes in natural vegetation, differences in seasonal and day/night appearance, and the presence of transient objects such as vehicles and people, all conspire to make recognition in its most general form a very hard problem.

Research into recognising places using vision has made progress, both in the robotics and in the computer vision communities [2–5]. Broadly speaking, approaches fall into two main categories: those based on matching local features between views; and those based on comparing whole image characteristics. Of the former, techniques based around the scale-invariant feature transform (SIFT) [6] and its variants are the most common, whilst in the latter category the GIST



(a) (b)

Fig. 1. Place recognition using landmark distribution descriptors (LDDs). Proposal regions from *Edge Boxes* (top left) are represented by CNN feature vectors and stacked in horizontal position order into an LDD for the view (left middle). The similarity of top matching regions within sections of the descriptor are then used as a measure of similarity between the views (bottom left). The approach outperforms comparison methods over 10 datasets each with 200 urban locations (top right) and shows excellent discrimination characteristics as illustrated by the confusion matrix in the bottom right.

descriptor [7] has found widespread use. To aid robustness, these techniques are often incorporated within some form of temporal integration, the probabilistic FAB-MAP method [8, 9] being the most well-known. Other techniques aim to deal with seasonal, day/night and long term changes, see e.g. [2].

As pointed out in [2], the two categories above tend to address complementary issues: local features provide a degree of invariance to viewing position and direction, whilst global descriptors provide better invariance to changes in viewing conditions. However, neither do both. To address this, recent work, for example that described in [1] and [10], match local regions corresponding to salient landmarks in the scene such as buildings, trees, windows, etc. Matching these regions using global-type descriptors provides a degree of invariance to changing conditions, whilst their localised nature gives better invariance to viewing position and direction. We adopt a similar approach in this work.

1.1 Landmark distribution descriptors

Our main contribution is that in addition to matching landmark regions, we seek to maintain consistency of the spatial distribution of landmarks between views of a place. In doing so, we aim to reduce the impact of similar landmarks being present in different places - although individual landmarks may match, it

is their relative positions across the view that characterises the place. In this work we limit ourselves to cases in which the different views of a place contain the same panorama but viewed from a different angle and distance, so that to a reasonable approximation the order of the landmarks, from left to right, say, remains the same between views. This accounts for many recognition scenarios, in which places are approached from the same general direction.

To implement this, we characterise a place using a *landmark distribution descriptor* (LDD), which consists of landmark feature vectors stacked in horizontal position order. Comparison of these descriptors then imposes the constraint of maintaining landmark order alongside matching feature vectors. We find that comparison is best achieved by identifying closest landmark pairs within vertical sections of the panorama, corresponding to subsets of adjacent feature vectors in an LDD (we used 3 sections in the experiments), and summing up distances between the respective feature vectors. We ensure view coverage by requiring sufficient numbers of proposal landmarks within each panoramic section. An example is shown in Figure 1a.

For landmark regions and features we follow the same approach as Sünderhauf et al. [1] and use *Edge Boxes* [11] and convolutional neural network (CNN) features, specifically AlexNet [12], followed by Gaussian random projection [13] for dimensionality reduction. Our use of panoramic sections to ensure view coverage also mirrors the tiling approach adopted in [10], although it is important to note that landmark ordering was not used in that work.

To evaluate the approach, we carried out experiments using image pair datasets for places in urban environments. Each dataset consisted of 200 places, with one image pair per place taken from different viewing positions. We used Google Streetview and Bing Streetside images so that image pairs were captured at different times and in different conditions. Results demonstrate that for 10 datasets in 6 different cities, our method performs consistently and significantly better compared with that obtained using the method in [1] and whole image matching using SIFT, GIST and CNN features. For example, as shown in Figure 1b, using 25 region proposals per view, on average over all datasets our method yielded an increase in precision (for 100% recall) of around 12% over that using whole image CNN, 20% over that using the method in [1] and 25%-30% over that using whole image SIFT and GIST.

The paper is organised as follows. In the next section we provide an overview of the system and details of the implementation of each component. Details of the datasets, the experiments and an analysis of the results is then provided in Section 3. We conclude with an indication of future work.

2 System Overview

In common with other place recognition systems we pose the problem as one of matching different views of the same place taken at different times. Labelled views are assumed to be held in a reference database and the task is to determine the most likely place associated with a test view captured 'online'. In this work

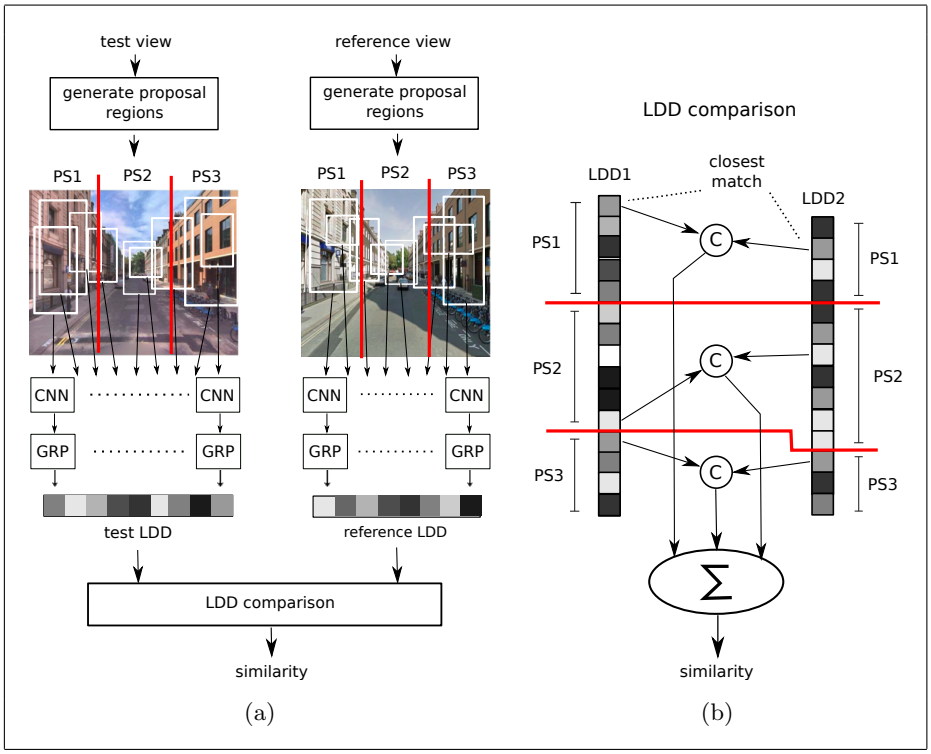


Fig. 2. Construction and comparison of landmark distribution descriptors (LDDs). (a) Landmark proposals are generated for the test and reference image using Edge Boxes[11], distributed within panoramic sections PS1-3; landmark features derived from a convolutional neural network (CNN) [12] followed by Gaussian random projection (GRP) [13] are then stacked in horizontal spatial order to form an LDD for each image; descriptors are then compared to derive a distance score between views. (b) Descriptors LDD1 and LDD2 are compared by identifying closest landmark features within each panoramic section and summing the (cosine) distances between them to derive an overall distance score.

we opt to consider the image-pair version of this framework, in which we have one reference view per place and successful recognition corresponds to matching the test view with the one correct reference view above all others. This contrasts with the majority of other evaluations, which have been based on matching frames within videos taken along the same route and successful recognition then being defined as matching a test frame in one video with a frame from a window of frames in another reference video. We discuss this further in Section 3.

Given the above, we now concentrate on how we match test and reference views. There are two components to this: constructing and comparing LDDs. These are described below and Figure 2 provides an illustration of the key elements of each.

2.1 Constructing LDDs

There are two main components to constructing an LDD for a given view as illustrated in Fig. 2a. First, proposal regions are generated, with the aim that a subset of these will correspond to salient landmarks. Second, feature vectors are computed for each of these regions, which are then combined into a single descriptor by stacking them in left-right position order.

Landmark proposals There are a number of algorithms available for generating proposal regions. In common with Sünderhauf et al. [1] we choose to use *Edge Boxes* as described in [11], which has found widespread use in object recognition and proved to be effective for our application. In brief, a valid edge box is identified as one in which there are a large number of contours wholly enclosed by the box. This is based on the observation that whole contours are likely to correspond to the boundary of distinct objects and hence that such boxes form good proposal regions suitable for further processing. This applies in our case as the landmarks we are interested in such as buildings, windows, trees, etc, satisfy this criterion. Also important is the fact that edge boxes can be found rapidly using fast edge detection combined with fast grouping of pixels into contours. We also make of the edge box ranking in [11] in order to limit the number of proposal landmarks and further speed up computation.

We are also interested in distributing landmark proposals across a view so that we can create a complete description. We do this by partitioning the image vertically and requiring that we select a fix number of the highest ranking landmark proposals in each section. We call these *panoramic sections* and in the experiments we used 3 sections: left, middle and right, such as that shown in Fig. 2a. In the main experiments these were positioned in a regular fashion about the image centre as shown but with overlap between sections to avoid excluding proposals which straddle a section boundary. The alternative is to align the sections according to the content of the view. We experimented with using the vanishing point (VP) as the centre and this proved effective for certain locations. We discuss this further in Section 3.

More formally, we denote by $L = \{l_1, l_2, \ldots, l_N\}$ the set of landmark proposals in an image discovered by the *Edge Boxes* algorithm. We then select a subset of landmarks $\hat{L}$ such that $\hat{L} \subset L$ and

$$\hat{L} = \bigcup_{s=1}^{S} \hat{L}_s \tag{1}$$

where $\hat{L}_s$ is a subset of top ranking proposals in panoramic section s and S is the number of sections, i.e. $S = 3$ in the experiments. We fix the number of top ranking proposals according to the section as described in Section 3 and deem a proposal to be in a section if its edge box is wholly within the section. Note that when using overlapping sections then individual landmarks can belong to two adjacent sections. This proves to be important when matching landmarks as it reduces the sensitivity to the positioning of section boundaries.

Landmark feature vectors To match landmarks between views we compute feature vectors to represent the appearance of the regions associated with landmarks. As illustrated in Fig. 2a, we again take the same approach as used in [1] and make use of convolutional neural network (CNN) features [12] followed by Gaussian random projection (GRP) [13] for feature vector size reduction.

CNN features have been shown to provide high levels of invariance to different lighting conditions and viewing positions [14] and hence are ideal for place recognition. Specifically, we used the pre-trained AlexNet network [12] as provided by MatConvNet [15] and extracted the feature vector of the 3rd convolutional layer (*conv3*). Landmark regions were resized to match the required network input size of 227×227 pixels and *conv3* produces feature vectors of dimension $13 \times 13 \times 384 = 64,896$.

To reduce the computational load when comparing feature vectors, we project each vector onto a lower dimensional space using GRP [13]. This is a simple but effective method for dimensionality reduction in which feature vectors are projected onto a significantly smaller number of orthogonal random vectors in such a way that with small error the distances between vectors is maintained. This makes it ideal when matching is based on comparing those distances as in our case. For the experiments we reduced dimensionality down to 1024 for each feature vector without significant impact on performance. In the GRP we used the integer based random projection matrix suggested in [16].

For a given view, we construct feature vectors for all the landmark regions in the selected subset of proposals $\hat{L}$ and the vectors corresponding to the section subsets $\hat{L}_s$ then form the LDD for the view. In the experiments we used a total of 25 or 50 proposals per view distributed over 3 panoramic sections and thus each descriptor was of size 25×1024 or 50×1024, respectively.

2.2 Comparing LDDs

For place recognition we seek the closest LDD within the database to that of the test image. To compare LDDs we could simply use the Euclidean distance between them. However, this assumes that we have successfully detected the same landmarks in each view, which is unlikely to be the case since we are generating proposals based purely on the appearance of each view individually, and not on the likelihood that a similar landmark exists in the matching view. Hence we transfer the latter constraint into the comparison process.

As illustrated in Fig. 2b, we do this by determining the best matching feature vectors (in terms of their cosine similarity) in each of the corresponding panoramic sections. Thus, for example, given two LDDs and using 3 panoramic sections, we seek the best matching pair in each section and then compute an overall matching score corresponding to the sum of the 3 cosine similarities between the feature vectors associated with each pair. We found that using cosine similarity, again in common with [1], gave improved performance over using a straight Euclidean distance.

More formally, given two descriptors, LDD1 and LDD2, containing landmarks

$$\{\hat{L}_1^k, \hat{L}_2^k, \ldots, \hat{L}_S^k\} \tag{2}$$

for $k = 1, 2$, we seek the set of S pairs $(\hat{l}_i^1, \hat{l}_j^2)^s$, $1 \leq s \leq S$, such that

$$(\hat{l}_i^1, \hat{l}_j^2)^s = \underset{l_i^1 \in \hat{L}_s^1, l_j^2 \in \hat{L}_s^2}{\arg\max} \; c(\mathbf{v}_i^1, \mathbf{v}_j^2) \tag{3}$$

where $\mathbf{v}_i^1$ and $\mathbf{v}_j^2$ are the feature vectors associated with landmarks l_i^1 and l_j^2, respectively, and $c(\mathbf{u}, \mathbf{v}) = \mathbf{u}.\mathbf{v}/||\mathbf{u}||||\mathbf{v}||$ denotes the cosine similarity between two vectors $\mathbf{u}$ and $\mathbf{v}$, where '.' denotes the dot product and $||\mathbf{u}||$ is the length of $\mathbf{u}$. To avoid duplicating matching landmarks, we also require that no landmark can be in more than one matching pair. The overall similarity score between the two LDDs, and hence the two views, is then given by the sum of the S cosine similarities, i.e.

$$sim_{12} = \sum_{\substack{(\hat{l}_i^1, \hat{l}_j^2)^s \\ 1 \leq s \leq S}} c(\hat{\mathbf{v}}_i^1, \hat{\mathbf{v}}_j^2) \tag{4}$$

3 Experiments

3.1 Datasets

We evaluated our method using multiple image pair datasets taken from urban environments. Our motivation for using image-pairs in contrast to matching frames in video sequences as used by others is twofold. First, we believe that it presents a more challenging test, since recognition is based on matching with only one alternative view as opposed to matching to one of multiple frames in a video (corresponding to the vicinity of a place). Secondly, it means that we can easily create large datasets corresponding to random places using the images taken from online mapping services such as Google Streetview, Bing Streetside and Mapillary. Using images from more than one of these also enables us to evaluate performance under different viewing conditions.

For the experiments reported here we used datasets obtained from Google Streetview and Bing Streetside. Specifically, we selected 200 random locations in 6 different cities - London, Bristol, Birmingham, Liverpool, Manchester and Paris - and for each location we selected images taken in roughly the same sort of direction but displaced by between approximately 5 and 10 meters. We selected one image from Streetview and one from Streetside for each location. This is ideal as the images were taken at different times and under different lighting and visibility conditions. We used datasets from different cities to enable us to evaluate the performance of the method for differing types and characteristics of architecture and urban layout. We tested the method on 10 datasets in all, using 3 from London and 3 from Bristol in order to test for any variation in performance within the same city. In total, the evaluation involved 2,000 different locations.

Example image pairs from the different cities are shown in Fig. 3. Note that although the physical distance between the viewing positions is not great, there

Fig. 3. Examples of view pairs from each of the 6 cities in the 10 datasets used in the experiments. The pairs are shown one above the other and there are 3 pairs per row.

is a significant change in structural appearance which when coupled with the differences caused by different light and visibility conditions makes recognition far from straightforward. Notable difficulties include the presence of pedestrians and vehicles, significant changes in scale of buildings, some buildings disappearing from view whilst others come into sharper focus, and so on. However, careful observation should reveal that distribution of the key visible landmarks is maintained across the two views. It is this characteristic that we aim to exploit in this work.

3.2 Comparison methods

We compared the performance of our method with four other methods: the CNN landmark matching method of Sünderhauf et al. [1]; whole image CNN matching

[17]; whole image SIFT matching; and whole image GIST matching. Relevant details for each method are given below.

CNN Landmark matching (CLM)
: As noted earlier, the primary difference between our method and that in [1] is that matching in the latter is based on finding similar landmarks across both views, irrespective of relative position. Specifically, best matching pairs of CNN-GRP feature vectors are selected from edge box proposals based on cosine similarity and the overall similarity between two views is then the sum of the cosine similarities, weighted by a measure of similarity in box size. For comparison, we evaluated two versions of this method, one using 25 (proposals CLM-25) and one using 50 proposals (CLM-50)[1]

CNN matching (CWI)
: In this method, we used the same CNN-GRP features vectors as in [1] and in our method, but comparison between views was based on a single feature vector computed for the whole image. Cosine similarity was again used as the comparison metric. The method is similar to that used in [17].

Dense SIFT matching (SIFT)
: For this method we used a dense keypoint version of matching SIFT descriptors across both views [6]. Specifically, we used the implementation as provided in the VLFeat library [18].

GIST matching (GIST)
: Finally, we compared our method with whole image GIST matching, based on the implementation provided by Oliva and Torralba as described in [7].

3.3 Results

We compared the performance of our method against that of the comparison methods for all 10 datasets. Each dataset contained 200 view pairs from different locations, with one view taken from Streetview and the other from Steetside. In each evaluation, we used all the Streetside images as test images and the Streetview images were used as the reference images. We used precision (P) and recall (R) to measure performance, defined as $P = tp/(tp + fp)$ and $R = tp/(tp + fn)$, where tp, fp and fn denote the number of true positives, false positives and false negatives, respectively. A true positive was recorded if the test image was matched with the reference image taken at the same location, a false positive was recorded if the test image was matched with a reference image taken at a different location, and a false negative was recorded if a test image was deemed not to match any of the reference images based on a threshold of the

[1] For clarity with respect to our experiments, we should note that we found that the similarity metric provided in [1] did not give good performance and so in the interests of fairness we used a modified version which gave significantly better performance. Specifically, we modified equations (2) and (3) in [1] to be $s_{ij} = 1 - (\frac{1}{2}(\frac{|w_i - w_j|}{max(w_i, w_j)} + \frac{|h_i - h_j|}{max(h_i, h_j)}))$ and $S_{ab} = \frac{1}{n_a \cdot n_b} \sum_{ij}(d_{ij} \cdot s_{ij}))$, respectively.

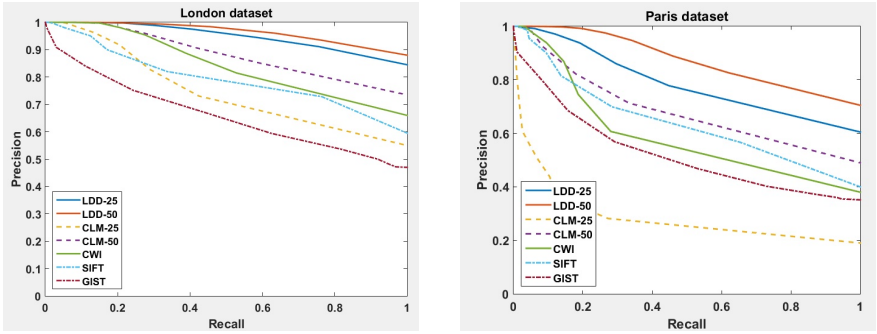
Table 1. Recorded precision values for 100% recall for all 10 datasets using all 7 comparison methods.

	LDD-25	**CLM-25**	**LDD-50**	**CLM-50**	**SIFT**	**GIST**	**CWI**
London1	**84.5**	55	**88**	73.5	59.5	47	66
London2	83	68	**90**	**84**	58	44	74
London3	**72**	57	**83**	69	51	58	64
Bristol1	**66.5**	51.5	**68.5**	58	51.5	33	60.5
Bristol2	**63.5**	50.5	**65.5**	59.5	40.5	26	54.5
Bristol3	59.5	47	**67**	**64.5**	48	37	61
Birmingham	**62**	44	**71.5**	60	26.5	38	44
Manchester	**69**	50.5	**71.5**	63.5	33.5	33.5	63
Liverpool	**74**	46	**75**	62	52.5	40.5	53
Paris	**61**	35	**70.5**	49	40	35	38
Avg	**69.5**	**50.45**	**75.05**	**64.3**	**46.1**	**39.2**	**57.8**

ratio between the closest and second closest matches. Variation of this threshold also enabled us to create precision-recall curves as given below. Note that our datasets do not contain any true negatives.

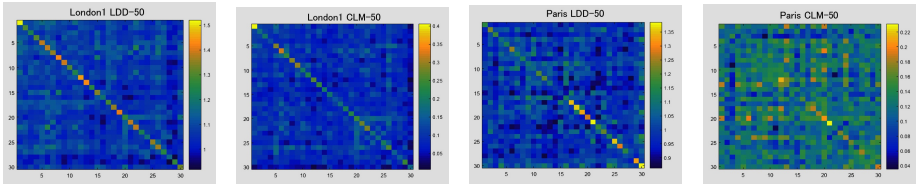
We evaluated two versions of our method, one using 25 landmark proposals and one using 50 landmark proposals. In each case we used 3 panoramic sections, with 50% overlap between sections. The image sizes were 640×480 pixels for both Streetview and Streetside and we used sections of size 320 pixels. We fixed the number of top ranked proposals selected from each section to be (5,15,5) when using 25 proposals (in left to right order) and (10,30,10) when using 50 proposals. The larger number of proposals in the central section proved to have a significant impact on performance.

Table 1 shows the precision values recorded for the different methods at 100% recall, i.e. so that all matches are accepted as positives. Note that over all datasets the LDD-50 method gives the best performance and apart from two datasets, the LDD-25 method gives the next best results. The latter is significant since the computational load is halved when using 25 proposal landmarks (the bottle neck is the computation of the CNN feature vectors) and thus it is interesting to note that good performance is still maintained from our method using the smaller number of proposals. This contrasts with the CLM method which performs significantly worse when using only 25 proposals and notably worse than using whole image CNN. We believe that this is a direct result of our method using the spatial distribution of the landmarks which provides a key characteristic to distinguish between views. Note also that the results for the London datasets are noticeably better than those for the other datasets. On inspection, we found that the London places were predominantly characterised by buildings with highly distinctive appearance, in contrast, for example, to the mix of vegetation and buildings apparent in the Bristol datasets and the similarities in architecture within the Paris dataset. This can be seen from the examples in Fig. 3. An area of future work will be to investigate how we can improve performance in such cases.



(a) (b)

Fig. 4. Precision recall curves obtained for all comparison methods for (a) the London1 dataset and (b) the Paris dataset.



(a) (b) (c) (d)

Fig. 5. Confusion matrices showing recorded similarity scores for 30 locations in the London1 and Paris datasets using (a)-(b) LDD-50 and (c)-(d) CLM.

Figure 4 shows the variation in precision as we reduce recall by increasing the number of false negatives via the threshold on the ratio of the closest and second closest matches for the two datasets London1 and Paris. Note that in both cases both versions of our method LDD-25 and LDD-50 outperform the other methods. Again, the difference in LDD-25 and CLM-25 is noticeable, with the former achieving almost a 30% gain in precision, corresponding to correct recognition of over 60 places compared with the latter, using the same number of proposal landmarks. This illustrates clearly the advantage of using landmark distribution to characterise views. To illustrate the distinguishing power of our method, Fig. 5 shows confusion matrices for the same two datasets using methods LDD-50 and CLM-50, where we have used 30 randomly selected location pairs rather than all 200 to aid clarity. These show the similarity scores between test and reference views. Note the high values down the main diagonal for the LDD-50 method indicating strong distinction of the correct places and contrast this with the closeness of the values obtained using CLM-50 method, especially for the Paris dataset.

To illustrate the landmarks that are being found by our method to enable correct recognition of places, Figures 6 shows examples of views which have been correctly matched. *None of these examples were correctly matched by the other methods.* In each case, the best matching landmarks found in each panoramic

Fig. 6. Examples of correct view matches obtained using the LDD-50 method. Matches are shown one above the other and there are 3 matches per row.

section are shown in colour, where the colours indicate corresponding landmarks in each view. Pairs are shown above one another and each row shows 3 pairs. Note the difference in appearance and structure between the views, especially the changes in vegetation and building structure, but also note that with careful observation they can be seen to be the same places. These are challenging examples and it is encouraging that our method is able to correctly match the views.

We also experimented with adapting the positioning of the panoramic sections according to view content rather than simply dividing up the image evenly into 3 sections about the image centre. Instead, we computed the location of the vanishing point in each view, using the method described in [19], and if within the image, we used this to centre the middle section, with appropriate adaptation of the two outer sections. In many cases this had little impact since the VP was often close to the image centre. However, in a number of cases it did make

Fig. 7. Use of the view vanishing point to center panoramic sections improves matching of landmarks (bottom) over that obtained using the image center (top).

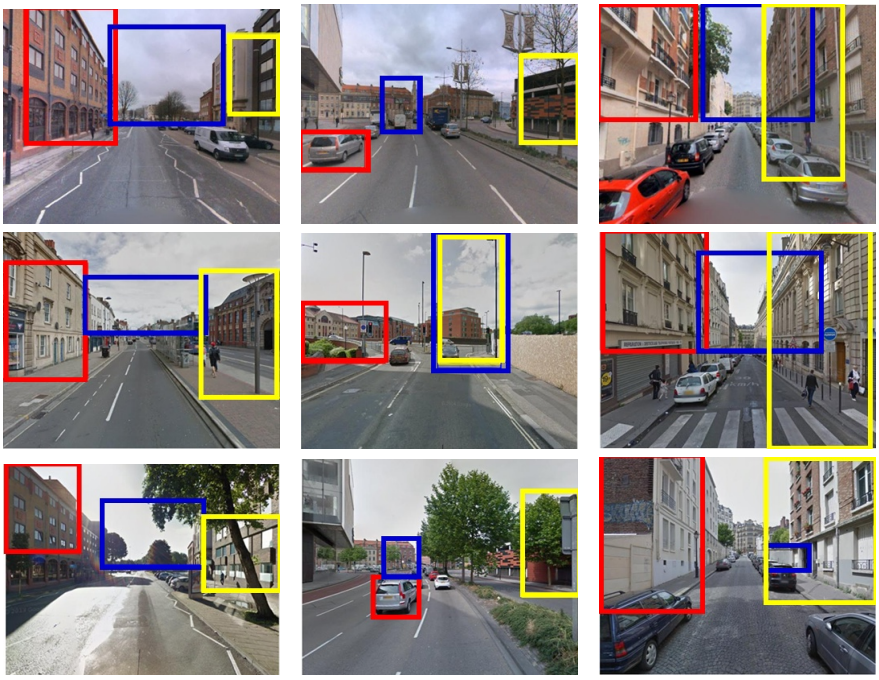


Fig. 8. Examples of incorrectly matched views obtained using the LDD-50 method.

a difference and resulted in correct matching of places which were previously incorrectly matched. An example is shown in Fig. 7. The top row shows a pair of views of the same place with selected landmark regions derived using panoramic sections centred about the image center. This proved not to be the best match for the left hand test image and hence resulted in an incorrect match. Clearly

the detected landmarks in each view do not correspond to the same landmarks in the scene. In contrast, shifting the sections to the right in both views after detecting the VP in each, results in correspondence between the detected landmarks and this resulted in a successful match. Although encouraging, these are only provisional results and further work is needed to determine the generality of using the VP in this way.

Finally, Fig. 8 shows 3 examples in which our method fails to match the correct view. The top row shows the test images, the middle row shows the incorrectly matched view and the bottom row the correct view. Note that these are particularly challenging examples and are further complicated by landmarks being detected on vehicles which are not present in both views. How to deal with cases such as these will be the subject of further research.

4 Conclusions and future work

We have presented a new method for visual place recognition based on matching landmark regions represented by CNN features. The key contribution is the encoding of relative spatial position of the landmarks via the use of the landmark distribution descriptors (LDD). Although the method has aspects in common with the CLM method of Sünderhauf et al. [1], we have demonstrated that the use of LDDs has a major impact on performance, with significant gains in precision, not only over CLM but also over the other whole image techniques. It is important to point out that the gains in precision amount to significant gains in the numbers of correctly recognised places, with, for example the 19% gain in average performance of LDD-25 over CLM-25 corresponding to 38 locations.

In the future we intend to investigate the performance of the method using different datasets, including video. We will also investigate further the benefits of using VPs to better position the panoramic sections. Also of interest is the potential for extending the idea of landmark distribution matching to more general cases in which landmark positioning changes due to changes in viewpoint. As the two are linked through geometry and motivated by the ideas and method described in [20], it may be possible to build this into a constraint for matching views which are widely disparate.

References

1. Sünderhauf, N., Shirazi, S., Jacobson, A., Dayoub, F., Pepperell, E., Upcroft, B., Milford, M.: Place recognition with convnet landmarks: Viewpoint-robust, condition-robust, training-free. In: Proceedings of Robotics: Science and Systems, Rome, Italy (2015)
2. Lowry, S., Sunderhauf, N., Newman, P., Leonard, J., Cox, D., Corke, P., Milford, M.: Visual place recognition: A survey. Robotics, IEEE Transactions on **PP** (2015) 1–19
3. Sattler, T., Leibe, B., Kobbelt, L.: Improving image-based localization by active correspondence search. In: European Conference on Computer Vision, Springer (2012) 752–765

4. Gronat, P., Obozinski, G., Sivic, J., Pajdla, T.: Learning and calibrating per-location classifiers for visual place recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2013) 907–914
5. Torii, A., Arandjelovic, R., Sivic, J., Okutomi, M., Pajdla, T.: 24/7 place recognition by view synthesis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 1808–1817
6. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision **60** (2004) 91–110
7. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. Int. J. Comput. Vision **42** (2001) 145–175
8. Cummins, M., Newman, P.: Fab-map: Probabilistic localization and mapping in the space of appearance. Int. J. Rob. Res. **27** (2008) 647–665
9. Cummins, M., Newman, P.: Appearance-only SLAM at large scale with FAB-MAP 2.0. The International Journal of Robotics Research (2010)
10. McManus, C., Upcroft, B., Newman, P.: Scene signatures: Localised and point-less features for localisation. In: Proceedings of Robotics Science and Systems (RSS), Berkeley, CA, USA (2014)
11. Zitnick, C.L., Dollár, P.: Edge boxes: Locating object proposals from edges. In: ECCV, European Conference on Computer Vision (2014)
12. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q., eds.: Advances in Neural Information Processing Systems 25. Curran Associates, Inc. (2012) 1097–1105
13. Bingham, E., Mannila, H.: Random projection in dimensionality reduction: Applications to image and text data. In: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '01, New York, NY, USA, ACM (2001) 245–250
14. Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A.: Return of the devil in the details: Delving deep into convolutional nets. In: Proceedings of the British Machine Vision Conference (BMVC). (2014)
15. Vedaldi, A., Lenc, K.: Matconvnet – convolutional neural networks for matlab. In: Proceeding of the ACM Int. Conf. on Multimedia. (2015)
16. Achlioptas, D.: Database-friendly random projections. In: Proceedings of the Twentieth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. PODS '01, New York, NY, USA, ACM (2001) 274–281
17. Sünderhauf, N., Shirazi, S., Dayoub, F., Upcroft, B., Milford, M.: On the performance of convnet features for place recognition. In: Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). (2015)
18. Vedaldi, A., Fulkerson, B.: VLFeat - an open and portable library of computer vision algorithms. In: ACM International Conference on Multimedia. (2010)
19. Kong, H., Audibert, J.Y., Ponce, J.: Vanishing point detection for road detection. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE (2009) 96–103
20. Frampton, R., Calway, A.: Place recognition from disparate views. In: Proceedings of the British Machine Vision Conference (BMVC). (2013)